# Improving the Performance of Recommender Systems by Alleviating the Data Sparsity and Cold Start Problems

**Guibing Guo**
Nanyang Technological University, Singapore
gguo1@e.ntu.edu.sg

## 1 Research Problems

Recommender systems, providing users with personalized recommendations from a plethora of choices, have been an important component for e-commerce applications to cope with the *information overload* problem. Collaborative filtering (CF) is a widely used technique to generate recommendations. The basic principle is that recommendations can be made according to the ratings of like-minded users. However, CF inherently suffers from two severe issues, which are the problems targeted in this research.

- *Data sparsity* refers to the difficulty in finding sufficient reliable similar users since in general the active users only rated a small portion of items;
- *Cold start* refers to the difficulty in generating accurate recommendations for the *cold* users who only rated a small number of items.

Lacking sufficient ratings will prevent CF from modeling user preference effectively and finding reliable similar users. In particular, the rating sparsity of recommender systems is usually up to 99% and cold users have rated less than five items in general [Guo *et al.*, 2012]. One of the resultant issues is the design of *similarity measure* to better model user correlation, especially in the *cold conditions* where only few ratings are present. To address these issues, two general strategies have been observed in the literature. The first strategy is to incorporate additional (user or item) information to help model user preference, and the second is to propose new similarity measures and make better use of existent user ratings. Although up to date many algorithms have been proposed, these issues have not been well addressed yet.

## 2 Progress to Date

We have conducted two different works following the general strategies mentioned in the last section. Specifically, we propose a trust-aware CF method to incorporate social trust[1] which is strongly and positively correlated with user similarity. In addition, we also propose a novel Bayesian similarity measure by taking into account both the direction and length of rating profiles whereas traditional methods only consider the direction of rating profiles.

### 2.1 Trust-aware Collaborative Filtering

The basic idea of our work [Guo *et al.*, 2012] is to form a new rating profile for the active users by merging the ratings of trusted neighbors, based on which accurate recommendations can be generated. During the merging process, three issues need to be taken care of. First, what kind of trusted neighbors will be involved. Since previous research shows that trusted neighbors may also share dissimilar preference, we only adopt users who have high similarity with the active users. Second, how to merge the rating of trusted neighbors. We opt to use the weighted average, i.e., for a certain item $i$, the ratings of trusted neighbors will be aggregated as follows:

$$\tilde{r}_{u,i} = \frac{\sum_{v \in TN_u} t_{u,v} r_{v,i}}{\sum_{v \in TN_u} t_{u,v}}, \tag{1}$$

where $r_{v,i}$ represents a rating reported by user $v$ and $\tilde{r}_{u,i}$ is the merged rating for the active user $u$. Note that user $u$ is also regarded as a trusted neighbor in the trust neighborhood $TN_u$, and $t_{u,v} \in [0,1]$ is the trustworthiness of user $v$ from $u$'s point of view (hence $t_{u,u} = 1$).

Third, whether we have sufficient confidence to adopt the merged rating. Two factors are taken into consideration: the number of ratings and the conflicts between positive and negative opinions. A rating is defined as *positive* if its value is greater than the median rating scale; otherwise, it is *negative*. The confidence of a merged rating $\tilde{r}_{u,i}$ is computed by:

$$c_{u,i} = c(r,s) = \frac{1}{2}\int_0^1 \left|\frac{x^r(1-x)^s}{\int_0^1 x^r(1-x)^s dx} - 1\right| dx, \tag{2}$$

where $r$ and $s$ are the number of positive and negative ratings used for merging process, respectively. Hence, the more ratings and the less conflicts between users opinions (for item $i$), the higher confidence we have to believe that the merged rating will be helpful. By repeating this procedure on the items rated by the trusted neighbors, a new rating profile can be generated for the active users.

Consequently, traditional CF can be applied to find similar users and make recommendations for the active users based on the newly formed rating profiles. Note that the rating confidence is taken into consideration when computing user similarity by Pearson correlation coefficient. We evaluate the effectiveness of our approach using three real-world data sets

[1] Trust reflects the extent to which users' opinions (ratings) are valuable in decision makings.

and compare with a batch of baseline methods. The experimental results show that our method generally and consistently achieves the best performance in terms of accuracy and coverage, especially in the cold conditions. In addition, we also analyze the capability of our approach in handling two extreme cold scenarios where (1) only trust is available; and (2) only ratings are available. Our method can survive in either case[2] and generate an effective rating profile for the cold users. Although it fails to function if neither trust nor ratings are present, this is beyond the scope of ratings-based CF. In summary, our approach shades light on a new way to take advantages of social trust and alleviate the concerned problems.

## 2.2 Bayesian Similarity Measure

Another line of research is to design new similarity measures, considering the ineffectiveness of traditional measures, namely Pearson correlation coefficient and cosine similarity. We propose a novel Bayesian similarity measure [Guo *et al.*, 2013] by taking into account both the direction and length of rating profiles (vectors), with the aims to solve the issues from which traditional approaches suffer.

The proposed similarity measure contains three components, namely *overall similarity* ($s'_{u,v}$), *chance correlation* ($s''_{u,v}$) and *user bias* ($\delta$). Formally, the user similarity between users $u$ and $v$ is computed by removing the chance correlation and user bias from the overall similarity:

$$s_{u,v} = \max(s'_{u,v} - s''_{u,v} - \delta, 0). \tag{3}$$

Specifically, for the overall similarity, we adopt the Dirichlet distribution to accommodate the multinomial ratings, or *rating distances* to be exact. We posit that not all rating evidences (i.e., a pair of ratings on the same items from two users) are equally useful for similarity computation, and that realistic similarity should be based on the consistency of ratings. The rating consistency is determined by two factors: (1) the standard deviation of ratings on a specific item; and (2) the rating tendency of all users. Then the posterior probability of rating distances can be updated according to the observed rating evidences. Hence the *user distance* can be computed as the weighted average of rating distances and the importance weights which reflect the extent to which the expected posterior probability is greater than the priori probability. In other words, the more evidences falling in a certain rating distance, the more important that level of distance will be. Lastly, the overall similarity is computed by inversely normalizing user distance. The second component is chance correlation, arising from the small number of user ratings leading to high chance to be 'similar'. The chance correlation is defined as the probability that the amount of evidences fall in different distance levels independently. In addition, by investigating the nature of traditional approaches as well as ours, we note that our method will generally hold a limited (yet much smaller) user bias, i.e., $\delta = 0.04$. We empirically demonstrate that removing chance correlation and user bias is helpful for similarity computation.

We exemplify the differences of the similarity values computed by traditional and our methods, and show that our method can solve the issues of traditional ones. More generally, we analyze the nature of similarity measures in terms of similarity mean and standard deviation. It is concluded that our method can achieve more realistic and distinguishable similarity measurements than traditional measures. Furthermore, we compare our method with a number of counterparts in terms of the accuracy of recommendations based on six real-world data sets. The results demonstrate that our method outperforms the others consistently.

# 3 Future Research

One potential drawback of our trust-aware recommender systems lies in its dependency on explicit social trust information. However, in real life people may not be willing to share their trust information due to the considerations of such as privacy. To address this issue, we would like to infer user trust information from the social interactions. For Bayesian similarity measure, we intend to incorporate more information about ratings, such as the time when ratings are issued, into the Dirichlet model in order to better model and capture the dynamics of user preferences.

Another direction that we have not studied yet is the dimension reduction methods. Clustering-based approaches will be our main focus in the future, considering their capability of involving new ratings as opposite to the static rating model (hence cannot adopt new ratings) built by matrix factorization methods. That also explains why we adopt a memory-based approach for our first work. However, literature shows that although clustering methods can benefit in scalability and efficiency, the general accuracy and coverage are relatively low. On the other hand, recent research [Bellogín and Parapar, 2012] has indicated that superior accuracy can be achieved if more sophisticated clustering methods are applied. For the coverage, our preliminary analysis is that by enabling the rating predictions for the cold users (the cluster of who is difficult to be identified due to little rating information), the coverage can be improved. Consequently, the accuracy may be further enhanced as well. As a natural expansion of our trust-aware recommender systems, trust-based clustering methods may take advantages of both trust utility and clustering capabilities, and further alleviate the data sparsity and cold start problems in recommender systems.

# References

[Bellogín and Parapar, 2012] A. Bellogín and J. Parapar. Using graph partitioning techniques for neighbour selection in user-based collaborative filtering. In *Proceedings of the 6th ACM Conference on Recommender Systems (RecSys'12)*, pages 213–216, 2012.

[Guo *et al.*, 2012] G. Guo, J. Zhang, and D. Thalmann. A simple but effective method to incorporate trusted neighbors in recommender systems. In *Proceedings of the 20th International Conference on User Modeling, Adaptation and Personalization (UMAP'12)*, pages 114–125, 2012.

[Guo *et al.*, 2013] G. Guo, J. Zhang, and N. Yorke-Smith. A novel bayesian similarity measure for recommender systems. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI'13)*, 2013.

[2] In the second case, our method will be equivalent with traditional CF since the active users are their only trusted neighbors.